The University of Reading

Met Office

Numerical Weather Prediction

Issues in high resolution limited area data assimilation for quantitative precipitation forecasting

Forecasting Research Technical Report No. 420
Joint Centre for Mesoscale Meteorology Report No. 139

Sarah Dance[1]

1 Department of Meteorology, University of Reading, Reading, UK

email: nwp_publications@metoffice.com

©Crown Copyright

ISSUES IN HIGH RESOLUTION LIMITED AREA DATA ASSIMILATION FOR QUANTITATIVE PRECIPITATION FORECASTING

S. L. DANCE

Department of Meteorology, University of Reading, PO Box 243, Earley Gate, Reading RG6 6BB

E-mail address: s.l.dance@reading.ac.uk

Date: February 2003.

Abstract. High resolution limited area data assimilation is a key aspect of improved quantitative precipitation forecasting. With advances in computer power it is now possible to envision operational models running at resolutions of O(1 km). To increase forecast skill, improvements in dynamics must be accompanied by improvements in initial conditions. Observations, such as radar winds and reflectivities, have the potential to provide enough high resolution information for good mesoscale analyses. A number of questions have to be addressed before this can become a reality. We review a number of issues in high resolution limited area data assimilation, including background and observation errors, balance, scale disparities, position errors and lateral boundary conditions. We go on to consider various methods for assimilation at high resolution and their relative advantages and disadvantages. In the operational context, the high computational cost of each method must be weighed against forecast improvements. More research is needed to determine the best method.

1. Introduction – Why do we need to assimilate precipitation data on the mesoscale?

Improved quantitative precipitation forecasting (QPF) is an important challenge for numerical weather prediction today. QPF presents potential benefits for a large section of society. In some cases, a better advanced knowledge of rainfall could save lives and livelihoods.

Mesoscale weather forecasts are typically generated from the output of a limited area numerical weather prediction (NWP) model. The inputs into these models are time dependent boundary conditions, usually provided from a larger scale model, such as a global NWP model; and initial conditions, based on an optimal combination of information from observations and previous forecasts. In this review we concentrate on the specification of initial conditions, also known as analyses.

In mesoscale NWP, a good representation of the initial state of the atmosphere is important for (Golding, 1987)

(1) Specifying forcing.
(2) Predicting the triggering and development of mesoscale disturbances.
(3) Predicting the decay of a disturbance once the forcing ceases.

Many mesoscale disturbances are forced either by the large scale flow or by the lower boundary. Thus it is clear that the forcing must be correctly specified in order to make a good forecast. A good description of synoptic scale flow requires a

Sarah Dance is supported by NERC Centres for Atmospheric Science.

consistent specification of the initial conditions for model variables in the same way as for a larger scale model. There are a number of contributing factors to the lower boundary forcing. Some types of lower boundary forcing e.g. orographic height are fixed, and could be specified to any required degree of accuracy in principle. However, an accurate three-dimensional representation of non-stationary atmospheric variables is required to represent other types of forcing, such as the thermal forcing occurring at the edge of thick cloud sheet or lying snow.

In many of these forced situations, the primary response of the atmosphere appears to occur through condensation of water and and consequent latent heat release. For example, the effects of latent heating play a central role in the development of convective systems. The dynamics of convection are nonlinear and have a very short range of predictability. Thus, in order to predict the mesoscale response to forcing, it would seem essential to have an accurate description of thermodynamic variables such as atmospheric moisture, heating, vertical velocity, potential temperature etc. in the analysis.

It may take several hours for a mesocale disturbance to develop once the forcing starts. Once the forcing ceases, the decay of the disturbance should also be predicted. For example, consider the case of an overnight forecast of a convective system. The overnight reduction in solar heating means that the system is unlikely to develop further. Thus a model initialised after sunset without the system present is unlikely to have very much skill.

Unfortunately it is difficult to generate an accurate mesoscale representation of the atmospheric state based on traditional synoptic observations. The details of the moisture field in particular are typically poorly represented in analyses based on such observations (Ducrocq *et al.*, 2000). This is especially important in the first few hours of forecast when precipitation forecasts suffer from a spin-up effect (balancing the hydrological cycle). This is caused by a combination of shortcomings in the subgridscale moist parameterisation and initial conditions.

Measurements of precipitation at fine scales can be accomplished using technology such as ground-based radar, microwave links and various space-based instruments. Such measurements are usually nonlinearly related to model variables. Incorporating such observations into the analysis presents new challenges. Whilst this is clearly an important part of the work that needs to be done, in this review we concentrate more on general issues related to high resolution data assimilation, rather than considering different observation operators (also known as forward models) for different kinds of observations.

With increases in computer power it will soon be possible to run operational mesoscale forecasts at O(1 km) resolution. This will give scope for modellers to begin to move towards explicit calculations of convection, rather than employing parameterisations. The hope is that the improvement in the representation of the dynamics will yield more accurate forecasts, but this cannot be accomplished without an accompanying improvement in analyses.

Before we can set about improving analyses, it is important to know what we mean by the optimum analysis. Meteorological dynamics are scale dependent. Similarly, what we interpret as a *good* analysis can change with scale between global and mesoscale domains. We discuss meteorological and mathematical aspects of optimality in section 2. In section 3 we consider the particular problems and unanswered questions associated with high resolution limited area assimilation. We go on to to consider the currently available methods and their advantages and disadvantages in 4. We conclude in section 5. A list of acronyms and abbreviations is provided in Appendix B.

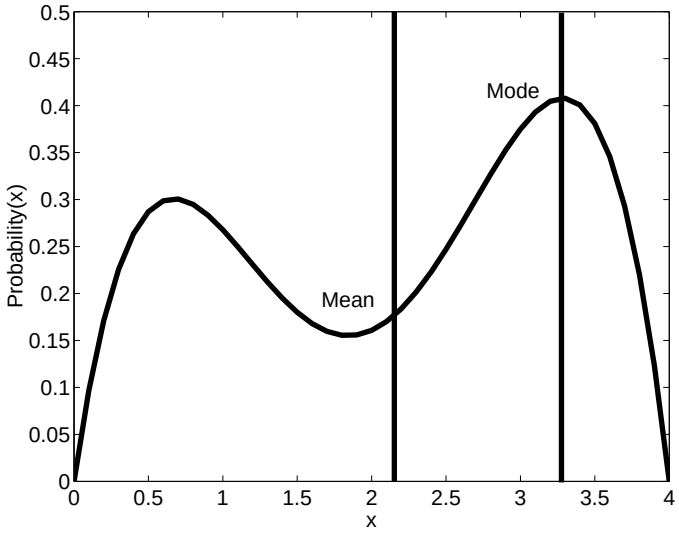


FIGURE 1. Schematic of a bimodal probability density function (pdf), with the mean and mode indicated using vertical lines.

2. What do we mean by the optimum analysis?

2.1. **Mathematical framework.** From a mathematical point of view, data assimilation has typically been approached from two standpoints: (1) A deterministic distance minimisation where we try to minimise the distance between a model trajectory and the observations in phase space. Here the distance could be the Euclidean norm, leading to the traditional least squares approach, or it could be any other metric. (2) A Bayesian probabilistic approach has also been used, where we consider the model state and observations as random variables, and the optimum analysis is chosen through knowledge of the conditional probability density function (pdf) of the atmosphere,

$$p(x, t|Y_t), \tag{1}$$

where x is the atmospheric state at time t, and Y_t represents all available observations.

The maximum likelihood estimate (MLE) of this probability distribution corresponds to the value of x which maximises the distribution. This is also known as the mode of the pdf. Choosing the MLE for the optimum analysis corresponds exactly to the solution of the least squares problem (Lorenc, 1986; Jazwinski, 1970).

The minimum variance estimate is another possible choice for the optimal solution. Under certain conditions, this corresponds to the mean of the distribution (Jazwinski, 1970).

Where the pdf is symmetric and unimodal (e.g. Gaussian distribution), the mean and the mode are equal, and the minimum variance estimate and the maximum likelihood estimate are exactly the same. However, in the more general case, for example a bimodal pdf, they will not be equal (see Fig. 1). What should be our choice of optimum in this instance? This is not just a theoretical question when it comes to data assimilation. For instance, if an observation being processed is close to being rejected as erroneous, the pdf of the corresponding atmospheric variable is bimodal (Lorenc, 2002*b*).

It may also be necessary to reconsider the role of the model.(Here the model we are referring to could be the full NWP model employed after the assimilation

cycle or the simplified model used in four dimensional assimilation.) We must ask the question, is the model designed to capture the evolution of the mean of the conditional pdf or the evolution of an individual trajectory in atmospheric phase space? With regard to precipitation forecasts, this may be especially pertinent. Moist parameterisations and turbulent closures are usually designed to be in some sense representative of an ensemble average. For example, the penetrative mass flux convection scheme used in the Met Office Unified Model, based on Gregory & Rowntree (1990), uses a parcel model based on the dynamics of small control volumes and ensemble averaged over different types of clouds. Model outputs for precipitation are averaged over a grid box, rather than representing point values of precipitation. It is important to note that spatial and temporal averages are not in general equivalent to calculating the mean of a conditional pdf at a point in space-time. With the arrival of higher resolution models and a move towards a lesser degree of parameterisation, we may have to reconsider the meaning of our output.

As a first approximation, data assimilation methods are often derived using Gaussian statistics and assuming a perfect linear forecast model. In this case it is known that the Kalman filter (KF) provides the optimal (minimum variance estimate) (Jazwinski, 1970). The full KF is not computationally feasible in meteorological systems, but many assimilation algorithms are related to the filter, so we describe it in 2.1.1 below. The reader may find it helpful to keep the KF equations in mind when considering later sections in this review.

In practice, error statistics are non-Gaussian (but not well known), models are imperfect and nonlinear. Some modern assimilation algorithms take these factors into account, but we cannot prove their optimality.

2.1.1. *Kalman Filter.* The discrete KF consists of difference equations for the evolution of the conditional mean and covariance matrix. We loosely adhere to the notation of Ide *et al.* (1997), but here we indicate the time-index and conditioning on previous times in parentheses.

Suppose we have a vector of unknowns, $\mathbf{x}(t_i)$ which evolves according to a model $\mathbf{M}$, such that

$$\mathbf{x}(t_{i+1}) = \mathbf{M}(t_i)\mathbf{x}(t_i). \tag{2}$$

For simplicity we have neglected model error, although it may be included in the KF formalism (Jazwinski, 1970). We write our best guess, or background forecast for $\mathbf{x}$ as $\mathbf{x}^b(t_i)$ and the corresponding background error covariance matrix as

$$\mathbf{B}(t_i) =< (\mathbf{x}^b(t_i) - \mathbf{x}^t(t_i))(\mathbf{x}^b(t_i) - \mathbf{x}^t(t_i))^T >, \tag{3}$$

where $\mathbf{x}^t(t_i)$ is the *true* solution at time t_i. As can be seen from equation (2), this evolves as

$$\mathbf{B}(t_{i+1}) = \mathbf{M}(t_i)\mathbf{B}(t_i)[\mathbf{M}(t_i)]^T, \tag{4}$$

assuming $\mathbf{M}$ is a linear operator and a perfect model.

Suppose we also have a set of discrete observations, related to the state-vector by a linear operator $\mathbf{H}$,

$$\mathbf{y}(t_i) = \mathbf{H}(t_i)\mathbf{x}(t_i) + \epsilon(t_i), \tag{5}$$

where $\epsilon(t_i)$ is a Gaussian white noise process with error covariance matrix $\mathbf{R}(t_i)$. In the meteorological context, we also assume that $\mathbf{R}$ accounts for the error of representativeness. This consists of contributions from the scales of the true atmosphere sampled by the observations but not resolved by the model.

Our best estimate or analysis for the state-vector, $\mathbf{x}^a$, given an observation $\mathbf{y}$ can be calculated from the algorithm illustrated in Fig. 2. In the algorithm, the error covariance of the analysis and background are also evolved with time. The operator that accomplishes this time integration, the matrix $\mathbf{K}$, is known as the

Initialisation:

$$\mathbf{K}(t_0) = \mathbf{B}(t_0)[\mathbf{H}(t_0)]^T\{\mathbf{H}(t_0)\mathbf{B}(t_0)[\mathbf{H}(t_0)]^T + \mathbf{R}(t_0)\}^{-1}, \quad (6)$$

$$\mathbf{x}^a(t_0) = \mathbf{x}^b(t_0) + \mathbf{K}(t_0)[\mathbf{y}(t_0) - \mathbf{H}(t_0)\mathbf{x}^b(t_0)]. \quad (7)$$

Analysis iteration: Loop over the time index, i

Step 1: Calculate the background field at t_i, using the analysis at t_{i-1}.

$$\mathbf{x}^b(t_i) = \mathbf{M}(t_{i-1})\mathbf{x}^a(t_{i-1}). \quad (8)$$

Step 2: Now calculate the background error covariance matrix at t_i.

- Calculate the conditional error covariance matrix associated with the analysis at time t_{i-1},

$$\mathbf{B}(t_{i-1}|y_{i-1}) = \mathbf{B}(t_{i-1}) - \mathbf{K}(t_{i-1})\mathbf{H}(t_{i-1})\mathbf{B}(t_{i-1}). \quad (9)$$

- Evolve this matrix under the model,

$$\mathbf{B}(t_i) = \mathbf{M}(t_{i-1})\mathbf{B}(t_{i-1})[\mathbf{M}(t_{i-1})]^T. \quad (10)$$

Step 3: Calculate the new Kalman gain matrix,

$$\mathbf{K}(t_i) = \mathbf{B}(t_i)[\mathbf{H}(t_i)]^T\{\mathbf{H}(t_i)\mathbf{B}(t_i)[\mathbf{H}(t_i)]^T + \mathbf{R}(t_i)\}^{-1}. \quad (11)$$

Step 4: Calculate the new analysis

$$\mathbf{x}^a(t_i) = \mathbf{x}^b(t_i) + \mathbf{K}(t_i)[\mathbf{y}(t_i) - \mathbf{H}(t_i)\mathbf{x}^b(t_i)]. \quad (12)$$

End loop over i

FIGURE 2. The discrete Kalman filter for a perfect model.

Kalman gain matrix. We have chosen to illustrate a simple version of the algorithm, where we assume a perfect, linear model. However, model error can be included, see Jazwinski (1970).

We now consider a very simple application of the filter in one dimension: estimation of a constant parameter by repeated measurements. Our intuition tells us that we would expect the optimal estimate to be the running average of all of these measurements: we will see that the KF provides a recursive method of calculating this mean, without storing all the previous measurements.

In the KF formalism, the constant parameter we wish to estimate is our one dimensional state-vector, x. The parameter is constant, so the model M describing the evolution of our state vector is just the identity, i.e. $x_{i+1} = 1.x_i$. Similarly the observation operator, H, relating measurements to the state vector, is also just equal to the scalar 1. We take an early measurement, y_{-1}, as our initial estimate of the background. We suppose that the measurements all have variance R, so the initial background error covariance, $B_0 = R$. Thus, the initial Kalman gain matrix is

$$K_0 = R[R + R]^{-1} = \frac{1}{2},$$

giving an optimal estimate, or analysis

$$x_0^a = x_0^b + K_0(y_0 - x_0^b) = \frac{y_{-1} + y_0}{2}.$$

For the next estimate, we set $x_1^b = x_0^a$. The corresponding background error covariance is given by

$$B_1 = MB(0|y_0)M^T = R - \frac{1}{2}R = \frac{1}{2}R$$

so the Kalman gain is $K_1 = 1/3$. Thus the new optimal estimate is

$$x_1^a = x_1^b + K_1(y_1 - x_1^b) = \frac{y_{-1} + y_0 + y_1}{3}.$$

Continuing to apply the filter, we see that indeed the recursion gives the mean as the optimal solution at each time.

More realistic systems have a much larger size. For example, a test configuration of the UK high resolution mesoscale model has 76 vertical levels and 300×300 horizontal grid points, giving a total of 6840000 grid points. Carrying out data assimilation using the KF with a state vector containing a only a subset of model variables, e.g. the four variables $\{u, v, \theta, p\}$ would lead to a state vector of size $O(10^7)$. The background error covariance matrix, $\mathbf{B}$ would then have $O(10^{14})$ elements. Even with a simple linear operator as the model $\mathbf{M}$, time evolution of $\mathbf{B}$ via a naive matrix multiplication would scale as $O(10^{14})$ operations. With a complex numerical weather prediction model for $\mathbf{M}$, calculating the evolution of this matrix becomes intractable. Thus the full KF is not operationally feasible.

Also note that the model in our example was assumed *perfect*. In the context of numerical weather prediction this assumption is patently false: all models have errors. Whilst the inclusion of model error in the KF formalism is a simple matter in theory, in practice this is difficult because the model errors are not known, so we must model the model errors. Additionally, it is hard to distinguish between model errors and initial condition errors, due to the cyclical nature of data assimilation algorithms, where previous forecasts are used as an a priori estimate for the next assimilation step. In operational data assimilation systems model errors are often neglected. A later section addresses this issue in more detail.

2.2. **Dynamical issues.** Our analysis is to be used as an initial condition for an NWP model. Thus the initial condition should be dynamically consistent (i.e. correspond to a possible model state). We must take care to avoid the presence of *unslaved* gravity modes in the initial condition which contaminate the solutions. (Note, however that we do wish to preserve *slaved* modes which are dynamically important at the mesoscale.) The process of adjusting analyses to avoid model spin up is known as initialisation. Some assimilation schemes incorporate constraints as part of the optimisation procedure so that the optimal analysis is already initialised. However, sometimes initialisation is applied as a post-processing step. This may move the analysis away from the mathematical optimum.

We would like to use as much information as possible from the observations in our analysis. In particular we would like to preserve any multivariate, spatial or temporal correlations which are also present in the true atmosphere. As a simple three dimensional example, consider using a pressure observation to generate a synoptic scale flow using the equations of geostrophic balance Daley (1991). In four dimensions, sequential observations of the same field can combine with the knowledge of the evolution equations to give more information about the dynamical fields (Lorenc, 2002*a*).

We note here that the word *balance* could be applied in both of these contexts: the avoidance of model spin up, and the preservation of certain relationships between variables in the analysis. As far as possible, we will make attempt to make this distinction clear.

2.3. **Verification.** We require an objective method of comparing data assimilation algorithms. As a first step we may test the fit of the data at analysis time. We can also look at the improvement in forecast skill using a numerical weather prediction model. For consistency, we should use the same criteria to judge the analysis and the forecast. Again, it is important here to bear in mind whether the model is designed to capture the evolution of the mean of the conditional pdf or the evolution of an individual trajectory in atmospheric phase space, and evaluate it accordingly.

Measurements of forecast skill are an active research area, especially for precipitation related variables. It is beyond the scope of this document to discuss this in detail. The interested reader is referred to the WWRP Verification Group website (Ebert, 2002) which provides an introduction and review of verification methodologies.

3. Challenges in mesoscale data assimilation

3.1. **Unknown background error pdfs.** Knowledge of the error pdfs enables optimal combination of inaccurate information. However, it is not possible to know the true state of the atmosphere at any given instant, so errors have to be estimated. These are usually expressed in the form of an error covariance matrix. The background error covariance matrix is defined by

$$\mathbf{B}(t_i) =< (\mathbf{x}^b(t_i) - \mathbf{x}^t(t_i))(\mathbf{x}^b(t_i) - \mathbf{x}^t(t_i))^T >, \tag{13}$$

(see section 2.1.1 for notation). Methods for estimating this matrix are discussed below.

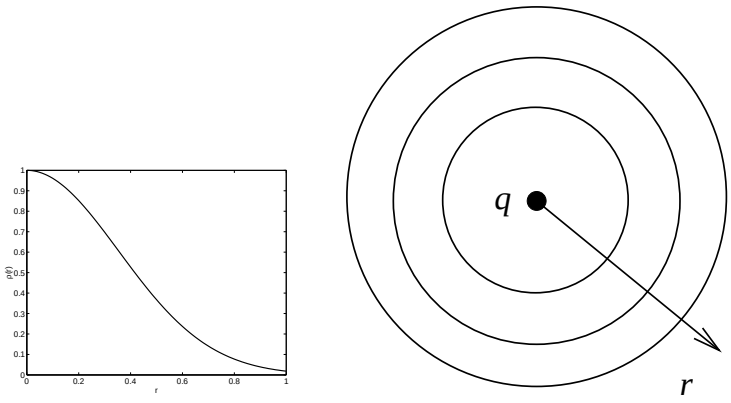
3.1.1. *The role of background error covariances in assimilation.* In analysis algorithms, background error correlations fulfil several roles (Bouttier & Courtier, 2002):

Information spreading: The shape of the analysis increment is determined by the analysis covariance structures. Observations are usually assumed to be spatially uncorrelated, thus it is correlations in **B** that spread the data in a finite region around each observation.

For example, consider a scenario with a single observation at position q, such that the observation operator is a the row matrix $H_j = \delta_{qj}$ and R is a scalar equal to the variance σ_R^2, of the measurement. Suppose the background error variance is equal to σ_B^2 and the background error correlations are isotropic. Say, $B_{iq} = \sigma_B^2 \rho(r_{iq})$, with r_{iq} the distance between points i and q. The Kalman gain for the system is easily calculated as $\rho(r)/(1 + \sigma_R^2/\sigma_B^2)$. This situation is illustrated in Fig. 3.

Information smoothing: Similarly, the amount of spatial smoothing of observations is determined by the correlation structure of **B**. The smoothing of increments is important in ensuring that the analysis contains scales that are statistically compatible with the physical fields. For example in low level mesoscale analysis it is important that structures triggered by orography are not smoothed out. At higher levels we would expect flows to be smoother.

Preserving error correlations: Nonlinear dynamics ensure that forecast errors project strongly onto a reduced dimension manifold (see section 3.1.2 below). The structure of the **B** matrix ensures that error correlations are preserved and that the analysis lies on this manifold. Some of these error



(a) (b)

FIGURE 3. Schematic to illustrate the spreading of observation data by the background error correlations. (a) Horizontally isotropic background error correlation function ρ as a function of radial distance r. (b) Contours of the analysis increments about a single observation located at q with a constant background field, and isotropic background error correlations, $\rho(r)$. See section 3.1.1 in the text for details.

structures may be flow dependent e.g. in an area prone to cyclogenesis one expects the most probable background errors to have the shape of the most unstable structures. Ideally the matrix $\mathbf{B}$ should reflect this flow dependence.

3.1.2. *Error estimation.* Where there is a dense observation network available, it is possible to obtain empirical estimates of background errors by calculating histograms of background error departure covariances stratified against separation. This method was popularised in meteorology by Hollingsworth & Lonnberg (1986). It is clearly impractical for data sparse regions.

The most commonly used method for gathering statistics is the NMC method (Parrish & Derber, 1992). This characterises the background error matrix $\mathbf{B}$ by accumulating statistics of the differences between different lead time forecasts, with validity at the same time. For example, in a global model, we might choose to compare a series of day $n-1$, T+48 forecasts with day n, T+24 forecasts. For a high resolution model, the best choice of lead time difference would have to be determined. In order to obtain smooth statistics, many pairs of forecast differences must be used, In consequence, the model of the background error cannot be considered flow dependent.

In principle, the time evolution of the conditional background error pdfs can be calculated using Kushner's equation (Jazwinski, 1970). Where there are no observations, the equation reduces to the Fokker-Planck equation. In general, the solution depends on all moments of the pdf. In the case that the dynamics are linear and errors are Gaussian, the full pdf can be described using only the mean and the variance and the system can be solved exactly (cf. the Kalman filter). We have already seen (section 2.1.1) that even with a linear model, the size of meteorological systems would make solving this equation intractable. However, the Ensemble Kalman filter (EnKF, to be discussed below), uses an ensemble of forecasts to

generate background error statistics, which are evolved in time using a technique which approximates the Kalman filter. Additional techniques for augmenting a constant **B** matrix with a flow dependent part representing the fastest growing errors are a subject of active research e.g . the error breeding system at the Met Office (Adrian Semple, personal communication).

Since the forecast model equations are nonlinear, we do know that a Gaussian pdf initial condition does not remain Gaussian (Smith *et al.*, 1999). In particular, nonlinear forecast dynamics quickly organise errors so that they project significantly onto a reduced-dimension manifold (Daley, 1991; Hamill, 2002), which on global scales is known to be dominated by balanced errors. On the mesoscale, the dynamical balances may be different (see below), but if we assume mesoscale variability is bounded, there must also be a reduced-dimension manifold at this scale.

The error covariance matrix **B** is typically large, full and poorly conditioned. A number of error estimation methods seek to take advantage of this concept of a reduced dimension manifold, by using dynamical or statistical balances to reduce the size of the matrix (see section 4.5). We can entirely avoid storing the matrix **B** by using recursive filters (also discussed in section 4.5).

3.2. **Observation error pdfs.** Observation error variances are mainly specified according to a knowledge of instrumental characteristics. These can be estimated by using collocated observations, for instance.

In most data assimilation systems, observations are treated as temporally and spatially uncorrelated. This is not always a good assumption, for example, radar data (Alberoni *et al.*, 2000). There are statistical methods for dealing with temporally correlated observations (Jazwinski, 1970) that involve augmenting the model state vector. However given the numbers of such observations in meteorological systems, this could prove to have prohibitive computational cost. To deal with spatial correlations, the assimilation scheme should include structure functions for the observation errors. How to specify these error correlations is an open question. Additionally, most assimilation schemes are not designed to easily handle such errors, although it is theoretically possible. Usually these problems are avoided by careful data selection and preprocessing.

3.3. **Data pre-processing and quality control.** We have already noted that observations using technologies such as radar and satellite imaging are having an important impact on QPF. Such observations require pre-processing to remove errors and to ensure that the resolution of the data is comparable to that of the analysis. For example, raw radar data suffers from problems with data folding, clutter contamination, anaprop, attenuation and range effects. Alberoni *et al.* (2000) review radar data preprocessing and quality control and note that the treatment of errors in assimilation of radar data is at a relatively immature stage. Despite these problems, radar reflectivities are used operationally in the Met Office nowcasting system to good effect.

3.4. **Dynamical Constraints.** At global scales there are well known balances which hold in the mid-latitudes (e.g. geostrophy). On the scales of high resolution limited area models, less is known about the balances which hold. Indeed, modes which we regard as spurious in global models may be dynamically important at the mesoscale. Balances also depend on the flow regime. For example, in a developing storm, we expect all scales to be present, whereas triggering convection may be regarded in some cases as a balanced phenomenon strongly forced by orography. Thus the balances we choose to impose on our initial condition depend on the types of flow we wish to capture accurately.

The imposition of balance constraints as part of the optimisation algorithm producing the analysis remains an area of active research, and will be discussed in the context of particular assimilation methods below.

3.5. **Scale disparities.** It is well known that meteorological systems are multiscale phenomena. It is important to note that global scale forcing impacts on mesoscale behaviour and vice versa. As the dynamics is multiscale, we similarly expect the background errors to be scale dependent. Representing the background errors at different scales in a statistically consistent fashion represents a challenge to the data assimilation community.

Recall that the background errors are used to smooth and spread observational data: measurements of different types may have vastly different densities and be representative of flow structures of very different scales e.g. the UK has six operational radiosondes which provide information about wind speed and direction, temperature and moisture. Ascents are made twice a day, and measurements are made at sixteen standard vertical levels between 1000 and 10 hPa. In contrast a weather radar could provide reflectivities (from which we may infer rain) and doppler wind measurements with a time frequency of O(1 min), and a spatial resolution of about O(1 km) with a horizontal range of O(75 km) (subject to scanning strategies, attenuation etc.).

Sequential assimilation of observations (or observation batching) at the same resolution is a standard technique in Kalman filtering (Brown & Hwang, 1997), which will be discussed in section 4.4. To deal with scale dependencies one could go a step further and carry out a series of nested analyses on grids representative of the scale of the observations. A two-way nesting could have a positive impact on the large scale analysis also. However, Mitchell *et al.* (2002) found that in *perfect model* experiments, interpolation between grids generated model error and problems with model spin-up.

On the other hand, it may be important to carry out the analysis of all correlated observations simultaneously, rather than splitting observations into batches according to spatial scale. In this instance, one might have to explicitly include scale dependent penalties in the optimisation procedure.

Scale disparities are not limited to spatial dimensions, but also apply to time. For example, data from radar is available at O(1 min), but radiosondes only make ascents twice a day. Carrying out data updates at fine time resolution is computationally expensive, and must be balanced with improvements to the analysis.

As a specific example we consider radar wind retrievals. It is a major challenge to produce an analysis with smooth transitions between regions with and without radar data (Sun & Crook, 2001). There are several methods to proceed, these are reviewed in Alberoni (2000). The Weygandt *et al.* (2002*a*,*b*) method sequentially retrieves radar data, followed by all the other thermodynamics variables. Their algorithm consists of

(1) Wind retrieval
(2) Blending with background
(3) Variational adjustment of blended wind field to satisfy mass conservation and observed radial velocity
(4) Thermodynamic retrieval and moisture adjustment

A data update is carried out every 15 mins.

Applying such a method may avoid some of the difficulties associated with full model adjoint techniques, while still retaining some of their attributes e.g. use of dynamic constraints and time tendency information. However, there are only a

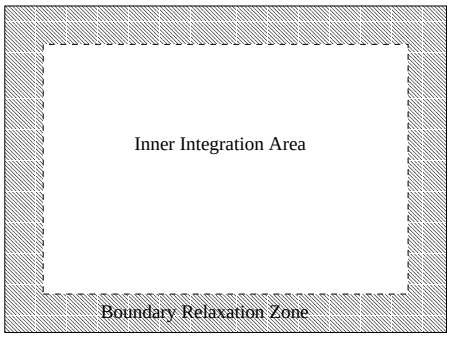


FIGURE 4. Schematic of inner integration area and relaxation zone for the LBCs for a full NWP LAM.

few case studies using this method published in the literature, and it is not clear whether such sequential systems will produce consistently good results.

3.6. **Lateral boundary conditions.** (LBCs)

Both the full forecast model and the data assimilation algorithm require lateral boundary conditions (LBCs). It is important to distinguish between these two separate requirements for boundary conditions.

First consider LBCs for limited area NWP models. Here the LBCs are usually provided from a larger scale or global model, and imposed through a relaxation condition, whereby the LAM forecast is nudged towards the global forecast in a relaxation zone around the boundaries (see Fig. 4). A discussion of such relaxation conditions is beyond the scope of this document. It is worth noting that the Met Office is unusual in that it has a *Unified Model* framework – the LAM is essentially a higher resolution version of the global model. More often, the global model used for the LBCs has a different formulation to that for the LAM e.g. the global model could be formulated in spectral space whilst the LAM could have a grid-based scheme.

Now consider the LBCs for the assimilation algorithm. In three dimensional assimilation, poor specification of the boundaries can influence nearby interior points due to smoothness and dynamical balance requirements. In four dimensional assimilation, errors introduced at the boundary may be advected across the domain and influence the interior.

The formulation of the assimilation algorithm may strongly influence which form of boundary conditions is best. For example, using EnKF the boundary conditions may be have to be supplied by a global ensemble; in 4DVAR, the dynamical model may be evolving increments rather than full fields, in which case the specification of the boundaries is not obvious. For example, we could simply set the increments to zero at the boundaries, or calculate the tangent linear model and adjoint for the full model boundary conditions and include the boundaries as control variables in the analysis.

3.7. **Data rejection.** The *effective time* for an observation is the time for which the observation can make a difference in a forecast. In limited area domains, weather systems advect out of the domain in a matter of hours, depending on the size of the domain. For example, in the UK mesoscale model the time period for information to be advected through the domain is about 18 hours (P. Clark, personal communication). Thus, even short effective times for observations can make large impacts.

For precipitation data, the effective time is typically only a few hours (e.g. Jones & Macpherson, 1997). This may be due to poor moisture parameterisations, but it may also depend to a certain extent on whether the model background state is close (in phase space) to the actual state of the atmosphere, i.e. whether the model state has correct synoptic forcing. In this case, we would like the data assimilation algorithm to shift the analysis away from the background state. Introducing rain into the analysis where there is no rain in the background has proved a particular problem for analysis schemes so far.

3.8. **Positional errors.** Sensitivity studies have shown that positional errors of precipitation data lead to significantly reduced forecast improvement, compared to rate errors (Jones & Macpherson, 1997). Position errors can also affect observation quality control. This is particularly problematic because small scale features (such as fronts, thunderstorms, or outflow in boundaries) may have been properly developed in the model, but have the incorrect location. We would like the assimilation system to remove incorrectly forecasted disturbances and rebuild them in the proper locations. This is especially challenging because such features are often temporally or spatially intermittent, accurate estimates of background error are lacking at such scales and the data may be scarce or incomplete.

A few authors have attempted to directly address location errors in the background field. Several of these methods require human intervention to identify the errors e.g. Alexander *et al.* (1998) use microwave integrated water vapour imagery and a digital warping or *morphing* technique, such as that employed in the film industry for special effects. This requires manually determined pairs of tie points to connect ares of similar features between the background and observed fields.

Automated systems have also been attempted e.g. The phase correction method of Brewster (2002) has been implemented in the ARPS system. Here a local translation is sought, described by a field of translation vectors $\delta\mathbf{x}$, that is to be applied to the background field in order to shift and distort it to match the observed data. The shift vector is found by minimising a weighted least squares difference between the background and radar reflectivity observations within a test volume. The shift vector is then applied equally to all prognostic variables, so that the structure and local balances between the variables will be nearly preserved, except for those containing local gradients. Rather than apply the shift in one step, the transition can be eased by applying the shift in small steps over the assimilation period, or by adding an artificial wind to blow the structures into their proper positions during the assimilation window. The phase shifting technique could serve as a preconditioning step before the application of other analysis schemes: phase correction does not correct the amplitude of the errors, so other methods will be necessary to complete the correction of the background forecast.

3.9. **Computational expense.** Although computer power is increasing rapidly, the choice of assimilation scheme for an operational system must still take account of the computational expense involved. Compromises may have to be made between choosing the best analysis system, and choosing a system that is operationally feasible. Consideration must be given to where analysis simplifications will make the least negative impact and the largest computational speed-up.

4. Available methods and their characteristics

4.1. **Physical Initialisation.** Physical Initialisation (also known as PI) was one of the first methods that enabled assimilation of precipitation data (Krishnamurti *et al.*, 1991). It was developed to make use of rain gauge and satellite outgoing longwave radiation (OLR) data to improve global forecasts in the tropics. Such

data do not correspond to prognostic model variables, but simulated measurements may be calculated from the model variables as diagnostics. The essential idea in PI is to work backwards through the model equations and calculate the values of the prognostic variables implied by the observations. The basic algorithm consists of two steps:

(1) A diagnostic calculation of surface fluxes and a subsequent humidity analysis consistent with the diagnosed fluxes, observed precipitation rates and observed OLR. The key concept behind the humidity analysis is inversion of the convective parameterisation scheme so that diagnosed precipitation rates correspond to those observed.

(2) Relaxation of the diagnosed fluxes into the model in a pre-integration nudging phase. This is accomplished using a Newtonian relaxation equation. Suppose A_i is a model variable satisfying the following evolution equation in the model:

$$\frac{\partial A_i}{\partial t} = F_i(\mathbf{A}, \mathbf{r}, t), \tag{14}$$

where $\mathbf{r}$ is the spatial coordinate. Then in the pre-integration phase we solve the nudging equation,

$$\frac{\partial A_i}{\partial t} = F_i(\mathbf{A}, \mathbf{r}, t) + \mathcal{N}(\mathbf{A}, \mathbf{r}, t)(\mathcal{A}_i^o - A_i), \tag{15}$$

where $\mathcal{A}_i^o$ is the diagnosed variable, calculated in step (1), consistent with the observed values. $\mathcal{N}$ is a penalty parameter, whose size determines how strongly we fit to the data. Such parameters may be different for each variable, and are determined by experience. For example, Krishnamurti *et al.* (1991) chooses each penalty parameter constant in time, with the penalty for the divergence equation much smaller than for pressure or vorticity, since the divergence field is strongly influenced by the imposed heating (prescribed rainfall rates).

PI does not guarantee initial conditions free of spurious gravity modes, so it is usually combined with a nonlinear normal mode initialisation scheme. For example this was the case in the old operational JMA mesoscale model, where radar/raingauge data was used for precipitation analyses. Non-precipitation variables were assimilated with a 3D-Optimal Interpolation scheme. Experiments showed that the method reduced both spin-up and position errors in precipitation forecasts. Aonashi (1993).(Note that the JMA has recently changed its mesoscale assimilation system to 4DVAR (Tsuyuki & Fujita, 2002).)

Several other mesoscale models make moisture and microphysics adjustments using the technique (see Ducrocq *et al.*, 2000, and references therein). Typically, the effective time of observations is a few hours. Shin & Ahn (2002) note that PI may result in excessive adjustment of analyses due to the imposition of strong constraints. PI is relatively computationally inexpensive and is fast enough to be used in an adaptive observations context (Shin & Ahn, 2002).

4.2. **Latent Heat Nudging.** Latent heat nudging (LHN) is a method motivated by physical ideas, whereby observations of rain-rate can be indirectly assimilated. Jones & Macpherson (1997) developed the scheme to incorporate radar data into an operational system (UK Met Office) which employed an analysis correction (AC) scheme for other variables. (The AC part of the system has been superseded by 3DVAR.)

The majority of water vapour which condenses in a cloud is precipitated out, and only a small amount of water and ice is stored in the cloud. Hence the rate of latent heating due to condensation in a control volume must be approximately

proportional to the net precipitation rate. In LHN, this idea is used to appropriately scale latent heat profiles to match more closely the observed precipitation rates. The particular implementation of Jones & Macpherson (1997) uses model profiles at each horizontal location where the model is wet. If we write the precipitation rate as PR, potential temperature as θ and use superscripts to indicate background and analysed values (as before), we have the following relationship for the potential temperature increments θ',

$$\frac{\theta^b + \theta'}{\theta^b} = \frac{PR^a}{PR^b}. \tag{16}$$

Before the increments are added to the model, they are smoothed by a recursive filter, to the resolvable scale of the model. The increments are then added to the model via Newtonian nudging or relaxation, in a pre-integration phase (see equation 15). This method avoids the need for explicit initialisation to reestablish dynamical and physical balances in the analysis.

Wang & Warner (1988) have studied the mechanism behind the forecast improvement for a convective case. This may be described in qualitative terms. The introduced imbalance in latent heat forcing provokes adjustment of the model's divergent circulation and vertical velocity field. In this case, the latent heat increments provide additional heating and the induced circulation gives more low level convergence. Additional moisture is entrained locally, decreasing the atmospheric stability. If LHN is employed together with another means of improving the moisture fields, then these circulation anomalies will tend to generate a condensation rate and hence precipitation rate more in balance with the applied heating. (LHN without moisture field adjustment is less effective.) More generally, the model's adjustment is influenced by the nature of the updraught sustaining the precipitating system, which can either be from large-scale or convective forcing.

The LHN method has been shown to give an increase in skill for the first 6-9 hours of forecast. Its main difficulty is encountered when the background forecast is dry, but precipitation is observed. This is dealt with by finding a nearby wet grid box and using the latent heating profile from there in order to introduce rain. However, this refinement has met with limited success. Care must also be taken to ensure that there is an upper and lower limit to the scaling factor on heating rates: without the former there would be a risk of feeding too much energy into the model, causing it to become unstable; without the latter one might remove too much heating, causing poor handling of heavy convective precipitation. The scheme does not involve inversion of any global matrices (where we use global in the sense of containing values from each grid-point in the domain), and so is relatively computationally inexpensive.

4.3. **Optimal Interpolation.** Optimal Interpolation (OI), (described in Daley, 1991), is derived as a minimum variance estimate for Gaussian statistics, and may be thought of as a simplification of the Kalman filter. It consists of carrying out an approximate inversion of the sum of the background, $\mathbf{B}$ and observation $\mathbf{R}$ error covariance matrices to derive interpolation weights $\mathbf{K}^{\mathrm{OI}}$. These are applied to the innovation vector of observation-background differences. This is given by

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{K}^{\mathrm{OI}}[\mathbf{y} - \mathbf{H}\mathbf{x}^b] \tag{17}$$

(cf. equation 12). As we have already seen (section 2.1.1), it is not possible to use the full Kalman filter since the matrices become too large to store and invert. The OI method relies on a hypothesis that for each model variable, only a few observations are important in obtaining the analysis increment. The essential idea is to restrict the Kalman filter calculation to involve only the relevant observations

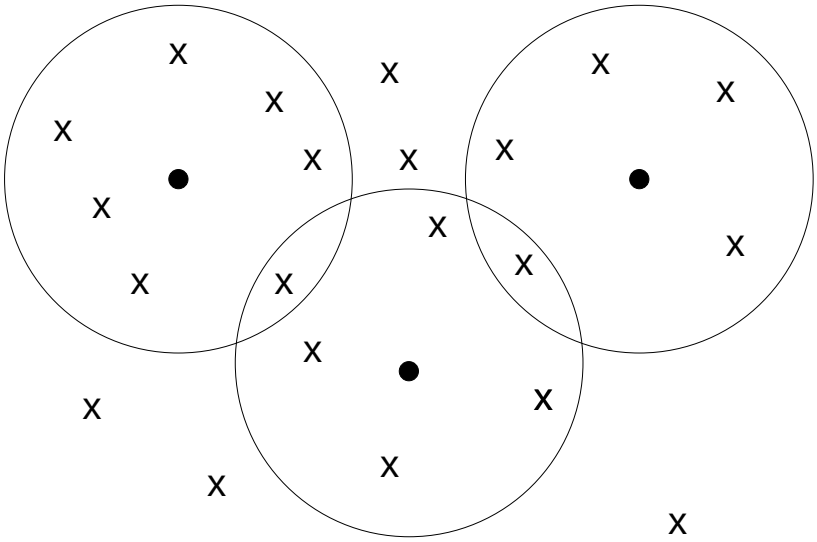


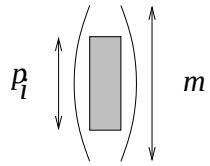
FIGURE 5. OI observation cut-off zones. Analysis points are marked with dots and observations with x. The large circles around each analysis point indicate the regions where the observational data is used for the OI calculation.

for each analysis point. The practical algorithm consists of (Bouttier & Courtier, 2002):

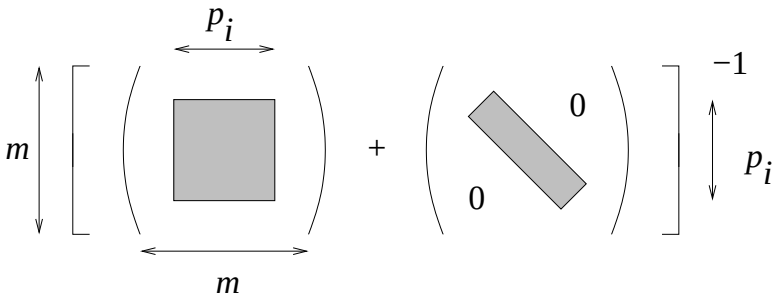
(1) For each model variable $\mathbf{x}(i)$, select a small number p_i of observations using an empirical selection criteria. The selection of observations should in theory provide all the observations which would have a significant weight in the optimal analysis. Practically speaking, usually we use a geometric distance based cut-off in order to determine which observations are used (see Fig. 5).
(2) Calculate the relevant list of p_i background departures $(\mathbf{y} - \mathbf{H}\mathbf{x}^b)$
(3) Calculate the $p_i \times p_i$ submatrix of $[\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R}]$ formed by restricting to the selected observations and invert this submatrix.
(4) Multiply by the relevant part of $\mathbf{B}\mathbf{H}^T$ (i.e. the i-th row restricted to the p_i observations) to obtain the i-th row of $\mathbf{K}^{OI}$.

We illustrate the resulting matrices schematically in Fig. 6 for a m observation, n analysis point system. Unless $\mathbf{H}\mathbf{B}\mathbf{H}^T$ is block diagonal, OI is not an exact equivalent to the Kalman filter. Local data selection can introduce spurious noise into the analysis fields (Lorenc, 1981). In particular, the set of observations used to perform the analysis at two neighbouring points may be different. Hence the analysis may not be continuous in space. The cost of the analysis increases with the size of selection domain, but compared to other methods it is simple to implement and computationally cheap.

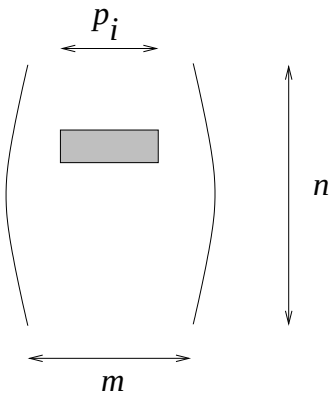
Nonlinear observation operators can be employed, but the underlying Kalman filter theory behind the method assumes linearity and a Gaussian distribution. Whilst the $\mathbf{B}$ matrix does not have to be globally defined for each variable, it must be locally positive definite, for the purposes of numerical inversion. This can also introduce noise, and makes it hard to guarantee a coherence of scales in the analysis (Lorenc, 1981). The method is only three dimensional and does not preserve temporal information.



(a)



(b)



(c)

FIGURE 6. Schematic to illustrate the OI approximation of the Kalman filter, from the full Kalman filter. The shading indicates the submatrices and vectors used in the approximation. (a) The restricted list of p_i departures, $\mathbf{y} - \mathbf{H}\mathbf{x}^b$. (b)$p_i \times p_i$ submatrix of $(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})$. (c) The i-th row restricted to the p_i observations of $\mathbf{B}\mathbf{H}^T$

OI has been widely used in operational systems, for example at NCEP in the Rapid Update Cycle (RUC) for radar winds. The system is described by Benjamin (1989). A number of other data assimilation methods are related to OI, e.g. the method of successive corrections and the analysis correction method previously used at the Met Office.

4.4. **Ensemble Kalman Filter.** We mentioned in section 2.1 that the Kalman filter provides the minimum variance estimate for systems described by Gaussian

pdfs and whose evolution is accomplished using a linear operator. NWP models are too large to use a full Kalman filter, but an ensemble approximation, the Ensemble Kalman Filter (EnKF), has been developed which could be feasible in an operational setting (e.g. Houtekamer & Mitchell, 2001).

The basic idea is to construct an ensemble of forecast states such that the mean of the ensemble is the best estimate of the population mean, and the sample covariance is a good estimate of the forecast error covariance of this best estimate. We denote this by

$$\mathbf{B}^f = \left\langle (\mathbf{x}^f - < \mathbf{x}^f >)(\mathbf{x}^f - < \mathbf{x}^f >)^T \right\rangle \tag{18}$$

where the angle brackets represent the ensemble mean over the forecast ensemble whose members are denoted by $\mathbf{x}_i^f$. The ensemble is integrated forward in time (using the full nonlinear forecast model if required), and the analysis $\mathbf{x}_i^a$ updated using the Kalman gain equation for each individual ensemble member (Lorenc, 2002*b*):

$$\mathbf{x}_i^a = \mathbf{x}_i^f + \mathbf{B}^f\mathbf{H}^T(\mathbf{H}\mathbf{B}^f\mathbf{H}^T + \mathbf{R})^{-1}(\mathbf{y} - \mathbf{H}(\mathbf{x}_i^f) + \varepsilon_i) \tag{19}$$

where ε is a noise process with zero mean and covariance $\mathbf{R}$. The best estimate is taken to be the ensemble mean of the analyses $\mathbf{x}_i^a$. The Ensemble Adjustment Kalman Filter (EAKF) is similar in many respects, but differs slightly in its method of updating the ensemble members (Anderson, 2001).

The analysis step can be applied sequentially to small batches of observations, with the analysed ensemble providing the background for the next batch of observations. As long as the errors in the observations are not correlated between batches, this is equivalent to treating them in one large batch. This is a standard technique used with the full Kalman filter: a proof of the equivalence between sequential and simultaneous processing is given in Appendix A.

Sequential processing is one aspect of the method that contributes to making it computationally feasible for large systems. The algorithm is also easily parallelisable, by distributing forecast runs over processors. However, every time a new batch of observations is assimilated, expensive matrix transpositions are required to be distributed over the processors. If frequent observations are assimilated, it is possible that communication costs may dominate.

The size of the ensemble can restrict the performance of the filter. The basic form of EnKF has only as many degrees of freedom as ensemble members to fit observations. One way to see this is by writing $\mathbf{B}^f$ as a product of two different matrices:

$$\mathbf{B}^f = \mathbf{X}\mathbf{X}^T \tag{20}$$

where $\mathbf{X}$ has as its columns the normalised deviations from the ensemble mean:

$$\mathbf{X} = \frac{1}{\sqrt{N-1}}(\mathbf{x}_1^f - < \mathbf{x}^f > \ \mathbf{x}_2^f - < \mathbf{x}^f > \cdots \mathbf{x}_N^f - < \mathbf{x}^f >). \tag{21}$$

$\mathbf{X}$ has column rank less than or equal to N, the number of ensemble members, and hence so must $\mathbf{B}^f$. Referring back to equation (19) we see that any increments to the background forecast must lie in the space spanned by the ensemble of forecast deviations (i.e. the columns of $\mathbf{X}$). In summary, if the background covariance structure cannot be represented accurately in a state space of size N, filtering methods are unlikely to work well without making use of other information about the covariance structure (Lermusiaux & Robinson, 1999).

One way of adding extra degrees of freedom to $\mathbf{B}$ is to use the Schur product (Houtekamer & Mitchell, 2001), discussed below. The EAKF (Anderson, 2001) uses a singular value decomposition and is well behaved if small singular values are omitted.

The initial ensemble is typically generated by adding noise to a reference trajectory. For small ensemble size, sampling errors in the covariance estimate can build up to give an ensemble whose covariance is too small. This is known as inbreeding. To counter this Anderson (2001) adds an extra covariance inflation factor, while Houtekamer & Mitchell (2001) split the ensemble in two, such that the analysis of one half is performed using the covariances estimated by the other.

Spurious covariances with distant grid points can do significant harm to the analysis. To avoid this Houtekamer & Mitchell (2001) use compactly supported covariances, avoiding discontinuities by using a Schur product (see below). However, local data selection and localisation of correlations can generate unslaved gravity modes. If we choose a sufficiently large radius, this problem can be overcome (Mitchell *et al.*, 2002). Moving to higher resolutions or more sophisticated models will require a larger ensemble, or more localisation (Mitchell *et al.*, 2002), so we may need to consider initialising analyses after applying the EnKF.

The forecast model used to integrate the ensemble members forward in time can be fully nonlinear. Nonlinear observation operators can also be implemented (Lorenc, 2002*b*). However, the use of nonlinear operators implies non-Gaussian pdfs, which the method does not allow for, resulting in suboptimal analyses.

EnKF relies on ensemble members evolving realistic error structures over an extended assimilation cycle. This may cause difficulties in limited area forecasting, where boundary conditions are specified from a global model. During the time taken for the EnKF to develop a realistic error structure, information from the boundary may have propagated across the entire area. Thus, the ensemble may be more influenced by errors from the boundary than from the analysis and error growth within the area. Perhaps a global ensemble would specify incoming errors from the boundary correctly. According to Lorenc (2002*b*) nothing has been published on this using ensembles.

4.4.1. *The Schur product in EnKF.* The Schur or Hadamard product of two matrices $\mathbf{P}$ and $\mathbf{Q}$ is the matrix $\mathbf{S}$ having the same dimension as $\mathbf{P}$ and $\mathbf{Q}$ such that

$$S_{ij} = P_{ij}Q_{ij}. \tag{22}$$

We denote this $\mathbf{P} \circ \mathbf{Q}$.

More specifically in the case of EnKF we carry out a Schur product of the background error covariance matrix with a matrix $A_{ij} = \rho(\|\mathbf{x}_i - \mathbf{x}_j\|)$, where ρ is a correlation function of local support, and $\mathbf{x}_i$ corresponds to row i of matrix $\mathbf{B}$ and $\mathbf{x}_j$ corresponds to column j. Thus the Kalman gain becomes

$$\mathbf{K} = [(\mathbf{A} \circ \mathbf{B})\mathbf{H}^T][\mathbf{H}(\mathbf{A} \circ \mathbf{B})\mathbf{H}^T + \mathbf{R}]^{-1} \tag{23}$$

Multiplication by the correlation function has several effects:

(1) It filters out small noisy correlations with remote points
(2) It improves the conditioning of $\mathbf{B}$ and allows the algorithm to function with even a modest number of ensemble members
(3) ρ is smooth and monotonically decreasing and so it tends to reduce and smooth the effect of observations at intermediate distances, resulting in smooth analysis increments.

4.5. **Variational Assimilation.** Variational assimilation consists of minimising a cost function composed of terms which penalise the distance between the analysis and the background error (J_b), and the analysis and the observations (J_o). In three-dimensions (3DVAR) we may write this as

Minimise

$$\begin{array}{rcccc} J(\mathbf{x}) & = & (\mathbf{x}-\mathbf{x}^b)^T\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}^b) & + & (\mathbf{H}(\mathbf{x})-\mathbf{y})^T\mathbf{R}^{-1}(\mathbf{H}(\mathbf{x})-\mathbf{y}) \\ & = & J_b & + & J_o \end{array} \tag{24}$$

where the notation is as before.

In space-time (four dimensions, 4DVAR) we seek an analysis $\mathbf{x}_0 = \mathbf{x}$ at initial time which will give a good fit to all the observations taken over the period of the assimilation window:

$$J(\mathbf{x}) = (\mathbf{x} - \mathbf{x}^b)^T \mathbf{B}^{-1} (\mathbf{x} - \mathbf{x}^b) + \sum_i (\mathbf{H}(\mathbf{x_i}) - \mathbf{y_i})^T {\mathbf{R_i}}^{-1} (\mathbf{H}(\mathbf{x_i}) - \mathbf{y_i}). \quad (25)$$

Here the subscript i corresponds to the sequence of observation times $t_0 < t_1 < \cdots < t_i < \cdots < t_n$. $\mathbf{x}_i$ must be a solution of the forecast model, M, i.e.

$$\mathbf{x}_i = M_{0 \to i} \mathbf{x}. \quad (26)$$

The variational formulation is related to the Bayesian probabilistic framework, in that the minimiser of the cost function is also the maximum likelihood estimate. The formulation given above is derived under the assumption of Gaussian statistics. Under this assumption, and supposing that the forecast model is perfect, the 4DVAR solution is equal to the solution obtained from the same observations and initial conditions using the Kalman Filter.

Cost functions corresponding to other non-Gaussian distributions correspond to non-Euclidean norms and may be derived analogously. For example, the JMA uses a non-Gaussian cost function for precipitation observation errors in their operational 4DVAR system (Kuizumi *et al.*, 2002). Additional terms may be added into the cost function as weak constraints, to ensure a smooth analysis, or to account for model error etc.

4.5.1. *Incremental formulation.* In operational models, the problem is formulated in terms of field increments, which may be at a lower resolution than the full field values. Given a full resolution guess solution $\mathbf{x}^g$ as well as the background $\mathbf{x}^b$, we seek a lower resolution perturbation model state $\delta\mathbf{x}$ that minimises

$$J(\mathbf{w}') = (\delta\mathbf{x} - \delta\mathbf{x}^b)^T \mathbf{B}^{-1} (\delta\mathbf{x} - \delta\mathbf{x}^b) + (H(x^g) + \mathbf{H}\delta\mathbf{x} - \mathbf{y})^T \mathbf{R}^{-1} (H(x^g) + \mathbf{H}\delta\mathbf{x}) \quad (27)$$

(for 3DVAR) where $\mathbf{B}$ is the covariance at the lower resolution. If we define a simplification operator $\mathbf{S}$ (possessing a generalised inverse, then $\mathbf{x}^a = \mathbf{x}^g + S^{-1}\delta\mathbf{x}$ and $\delta\mathbf{x}^b = S(\mathbf{x}^b) - S(\mathbf{x}^g)$. In 3DVAR usually the best guess estimate is taken equal to the background so $\delta\mathbf{x}^b$ is zero. In 4DVAR, as above, the cost function is modified to incorporate observation terms for a series of time steps. Here, the model $\mathbf{M}$ is in theory the tangent linear equivalent of that for the full formulation. However, in practice rather than compute the full tangent linear model of our NWP model (and its adjoint), we might design a perturbation forecast model to evolve the optimal analysis increment (i.e. the mean or mode of the analysis pdf, depending on your philosophy of optimality). Note the contrast to the EnKF forecast model that should evolve individual trajectories, which are then ensemble averaged to find the mean (Lorenc, 2002*a*,*b*).

The minimum of J is found numerically by an iterative method utilising the gradient of J, ∇J. This places certain requirements on the conditioning of the matrices $\mathbf{B}$ and $\mathbf{R}$. We calculate ∇J using the adjoint of the linearisation of the forecast model. Adjoints of sophisticated models can be difficult to code, if they exist at all. Computational and theoretical advantages and disadvantages of variational methods relating to mesoscale assimilation are highlighted below. 3DVAR and 4DVAR share many of the same characteristics, which will be discussed concurrently. Properties exclusive to 4DVAR are explicitly noted.

4.5.2. *Modelling the background error covariance matrix.* The background error covariance matrix is not known and must be modelled. In many operational systems, the NMC method (Parrish & Derber, 1992) is used. This characterises the matrix

B by accumulating statistics of the differences between different lead time forecasts, with validity at the same time. For example, in a global model, we might choose to compare a series of day $n-1$, T+48 forecasts with day n, T+24 forecasts. For a high resolution limited area model, the best choice of lead time difference would have to be determined, bearing in mind that the time period need to be long enough for errors to grow, but short enough that all the information has not been advected out of the domain.

Using the NMC method to model $\mathbf{B}$ results in a very large, full matrix which is typically ill conditioned. Hence transformations to a control variable space are often employed to reduce computational storage and expense in the minimisation. Ideally, the choice of control variables should be (Ingleby, 2002):

(1) **Normal.** The pdf for the variable should be normal, otherwise the cost function for the variational problem (24) should take a different form. Most atmospheric variables are likely to have non-normal error distributions due to the nonlinear nature of the dynamics (Smith *et al.*, 1999).
(2) **Uncorrelated.** This makes the $\mathbf{B}$ matrix diagonal, and decouples the system, enabling us to solve for each variable independently. In practice, we remove known dependencies between variables as far as is possible, by predicting one from the other using physical relationships or linear regression. We then treat the residual as the control variable.
(3) **Observable.** If the variable is not related to an observable quantity, the optimal analysed state is the background! Preferably the relationship between the observed quantity and the control variable should be linear.
(4) **Symmetric.** Translational symmetry implies that the variable is homogeneous, rotational symmetry implies that the variable is isotropic.
(5) **Local.**
(6) **Smooth.**

No choice of control variable satisfies all these properties in practice. As part of the transformation, dynamical balance constraints are often imposed (Parrish & Derber, 1992; Derber & Bouttier, 1999; Lorenc *et al.*, 2000; Gauthier *et al.*, 1999). It is not known if the inclusion of balance in this fashion is effective at small scales (Xue *et al.*, 2002). The way the balance equations are formulated (e.g. for geostrophic balance in terms of vorticity or rotational and divergent wind) may affect the scale at which the balance is imposed (Xie *et al.*, 2002).

In global models, the control variable is usually defined in horizontal spectral space using spherical harmonics. This is related to the possibility that different spectral forecast error modes are statistically independent, given the assumption that spatial covariances are horizontally homogeneous in physical space. A similar property exists in the spectral space of a limited area model. However, the application of spectral transforms over a limited model area poses a particular problem, since the fields do not have a periodic variation over the area. One possibility would be to assume zero-valued analysis increments on lateral boundaries and to apply sine transforms (as in Lorenc (1997)). The HIRLAM (Gustafsson *et al.*, 2001) uses an area extension to obtain periodic variations in both horizontal dimensions. This makes it possible to have non-zero increments on the lateral boundaries and to apply standard FFTs for the transformation. Gustafsson *et al.* (2001) note that forcing assimilation increments to be zero at the boundaries gives very unrealistic assimilation increments close to the boundaries.

Where there is a dense observation network available, it is possible to obtain empirical estimates of background and observation errors by calculating histograms of background error departure covariances stratified against separation. This method

was popularised in meteorology by Hollingsworth & Lonnberg (1986). It is clearly impractical for data sparse regions, and synoptically independent.

4.5.3. *Recursive filters as an alternative to storing* **B**. An alternative to storing the background matrix explicitly is the use of recursive filters, applied in physical space. There is an extensive literature on filters. An introductory review is given by Raymond & Garder (1991). As an example consider a one dimensional grid with spacing δx; A_i is the initial value at gridpoint i, B_i is the value at i after one pass of the filter, C_i is the value at the gridpoint after a pass of the filter in each direction and α is the filter coefficient. Then the basic recursive filter is defined by

$$B_i = \alpha B_{i-1} + (1-\alpha)A_i \text{ for } i = 1 \dots n \quad (28)$$

$$C_i = \alpha C_{i+1} + (1-\alpha)B_i \text{ for } i = n \dots 1. \quad (29)$$

This is a first order recursive filter applied in both directions to ensure zero phase change. Domain boundary conditions enable us to set B_0 and C_{n+1} appropriately, and α is chosen so that the filter gives a required spectral response.

Unlike the more common operational methods, of applying a covariance matrix in spectral space, recursive filters can allow for inhomogeneous covariances, but there is only limited freedom to specify the profile shape of the error statistics in wavenumber space (Wu & Purser, 2002). In practice the filters are often chosen to be Gaussian. Using a Gaussian shape can seriously hamper the ability of the analysis to assimilate the smallest scales. In particular, the power spectrum of a Gaussian shaped covariance is itself of Gaussian form, and drops off faster than the spectral decay of the real atmosphere. If we assume that for the smallest scales the background has little skill, then the error should have similar structure to the field itself, consequently the tails of the power spectra of the background error should be fatter than those of any purely Gaussian spectrum. A *fat-tailed* spectrum is easy to achieve in spectral space, but to achieve this using a filter requires use of a linear combination of recursive filters with different correlation scales.

In addition, the use of recursive filters has further problems: namely Gibbs phenomena due to filter truncation and complicated specification for non-uniform grids. (Huang, 2000).

4.5.4. *Flow dependent covariances.* On the mesoscale it may be especially important that the background error terms are flow dependent and represent the small scale structures in the flow accurately. This may be achieved using 4DVAR: although the background error covariance matrices are averaged over a long period, and the covariances not changed on a daily basis, there is an implicit dependence on the flow through the assimilation window, via use of the TLM or perturbation forecast model, see Lorenc (2002*a*).

4.5.5. *Observation of non-model variables.* A particular advantage of variational methods over many of the other methods we have discussed is that they allow direct use of observations related to the model variables via nonlinear observation operators. These are especially important in QPF e.g. GPS precipitable water, satellite radiances, radar radial velocity and reflectivity.

4.5.6. *Synergistic information.* Another benefit of 4DVAR is its ability to penalise solutions which are inconsistent with the dynamical equations when fitting the observations. It allows dynamically consistent information to be extracted from a time series of observations which would not be available from a three dimensional method (Lorenc, 2002*a*).

4.5.7. *Weak dynamical constraints.* We have already mentioned that dynamical balances can be imposed as strong constraints through the control variable transform. An alternative which could possibly be more appropriate on the mesoscale is to impose such balances as weak constraints. This has been tried in the Oklahoma ARPS system, where they impose mass continuity and a 3D diagnostic pressure equation as a weak constraint (Xue *et al.*, 2002; Gao *et al.*, 2002). They also have an option to penalise the difference between analysis and observed time tendencies of radial momentum. The use of a penalty method such as this does not spread error information among multiple variables in a physically consistent manner. In addition the proper choice of the size of the lagrange multipliers is not obvious. For an idealised case with the ARPS system, the analysed wind field is not sensitive to these parameters, but the thermodynamic field generally is (Jidong Gao, personal communication 2002).

4.5.8. *Smoothness penalty.* We noted earlier that for radar observations it is a major challenge to produce an analysis with smooth transitions between regions with and without data. Sun & Crook (2001) employ a spatial and temporal smoothness penalty to help smooth the analysis. It is given as

$$J_p = \sum_{t,i,j} \alpha_{1i} \left(\frac{\partial A_j}{\partial x_i} \right)^2 + \alpha_{2i} \left(\frac{\partial^2 A_j}{\partial x_i^2} \right)^2 + \alpha_{3i} \left(\frac{\partial A_j}{\partial t} \right)^2 + \alpha_{4i} \left(\frac{\partial^2 A_j}{\partial t^2} \right)^2 \tag{30}$$

where A_j represents any of the model dependent variables and x_i represents the spatial dimensions (x, y, z). Of course, instead of using first and second order derivatives we could use a different set of smoothing weights. In the temporal domain, the general penalty term is known as a digital filter.

The advantage of using a penalty term is that it does not depend on the quality of the background field. However, the weighting coefficients have to be determined from experience. Sun & Crook (2001) find that the analysis and subsequent forecast are not very sensitive to the magnitude of the weighting coefficients, within the range of a few orders of magnitude. This particular formulation only smoothes the analysis within a distance of a few grid points because it uses second order derivatives. In comparison with the penalty term, a background term with a good model of error statistics holds more promise, since the filtering performed through the use of appropriate error correlation (e.g. with a variable lengthscale) can provide more flexibility.

4.5.9. *Model error.* The use of the full model equations as a strong constraint in 4DVAR neglects model error (Weygandt *et al.*, 2002*a*). Zupanski (1997) accomplishes the inclusion of model error by imposing the model as a weak constraint:

$$\mathbf{x}_i = M_{0\rightarrow i}\delta\mathbf{x} + \eta_i. \tag{31}$$

where η is a noise process with covariance $\mathbf{Q}$. This results in the inclusion of an extra term in the cost function,

$$J_m = \sum_i \eta_i \mathbf{Q}_i^{-1} \eta_i \tag{32}$$

where the subscripts indicate the time index. We must also augment the control variable with the model error at each time step: $\delta\mathbf{x} = (\delta\mathbf{x}_0, \eta_0, \eta_1, \ldots)^T$ For LAMs, we could split model error into two parts, interior and boundary condition errors. Unfortunately, there are serious practical difficulties: the minimisation problem can become increasingly ill conditioned when the model error terms are considered control variables. Such problems can be overcome by decreasing the number of degrees of freedom for model error (coarsening in time).

Zupanski (1997) notes that the analysis and forecast results using the standard strong constraint formulation are quite different to those using model error as a weak constraint. Also note that since we do not know the pdf for the model error, we have to approximate it in the first instance and then calculate it from optimised statistics.

According to Zupanski (1997), the effect of including model error is in degrading the value of information in the distant past. Thus model error contributes more to the better fit of data at the end of the assimilation period than at the beginning. On the other hand, overweighting recent data in the strong constraint formulation would have a similar effect.

We note that since the model is no longer a strong constraint, the resulting analysis may require initialisation. Zupanski (1997) also uses a digital filter gravity wave penalty term in her formulation to prevent any associated unslaved modes.

4.5.10. *Adjoints of the tangent linear model.* To find the gradient of the cost function J it is necessary to integrate backwards in time using the adjoint of the linearisation or tangent linear model (TLM) of the forecast model. In principle this is straightforward for smoothly varying operators. However, moist parameterisations often contain on-off switches. In this case the tangent linear model and its adjoint are not well defined at the switch point. Several authors have considered this is more detail e.g. (Bao & Kuo, 1995). If the analogue of the switch function is continuous (but not necessarily differentiable) at switch time, the adjoint to the TLM is continuous at switch time. If however the switch is equivalent to a step function, calculation of the adjoint becomes complicated because sensitivity of the model state at switch time must be determined.

Switches are probably unphysical, and it would be desirable to develop parameterisations which do not use such switches. However, this is not the current state of affairs. As we move to higher resolution models, our dependence on parameterisations will decrease, and the hope is that switches will become less of an issue.

4.6. **Hybrid Methods – Variational methods using ensemble covariances.** Probably the weakest part of variational data assimilation methods is the specification of the background error covariances. These might be improved via use of ensemble methods. Ensemble covariances are degenerate, so we have to transform to a new control variable (Lorenc, 2002*b*). The method for generating the ensemble does not have to be EnKF, for example, we could use error breeding methods (Adrian Semple, personal communication).

5. Summary

High resolution data assimilation will be essential for improved QPF when computational capacity allows higher resolution operational forecasts. However, much work needs to be done before this can become a reality. We discussed a number of general issues relating to high resolution limited area analysis. A consideration of these issues clearly goes hand in hand with the choice of assimilation method. We identified a number of existing methods which may be suitable for a high resolution limited area system. Certain characteristics of each method are summarised in Table 1.

The table shows that there is no *best* method currently available: all have advantages and disadvantages. Future work will determine the optimal solution!

Category	PI	LHN	OI	EnKF	3DVAR	4DVAR
Nonlinear observation operators	No	No	No	No	Yes	Yes
Ability to fit many observations	Precipitation rate data only	Radar data only	Limited by matrix size: data fitted sequentially	Limited to fewer data than ensemble members in a given region	Limited only by resolution of analysis grid	Limited by resolution of simplified forecast model
Observation errors accounted for	No	No	Yes, with the assumption that the errors are Gaussian	Yes, with the assumption that the errors are Gaussian	Both Gaussian and non-Gaussian errors may be accounted for	Both Gaussian and non-Gaussian errors may be accounted for
Background errors accounted for	No	No	Yes	Yes	Yes	Yes
Flow dependent background errors	No	No	No	Yes	No	Yes
Model error accounted for	No	No	No	Yes	May be included	May be included
Lateral boundary conditions	??	??	??	May require a global ensemble	Set boundary increments to zero or treat as control variables	Set boundary increments to zero or treat as control variables

TABLE 1. Assimilation methods and their characteristics

Category	PI	LHN	OI	EnKF	3DVAR	4DVAR
Balanced increments	No	No	No	Balance may be lost in the Schur product	Yes	Yes
Uses forecast model	Inverts forecast model parameterization	Requires consistency with model parameterization - scales model output	No	Uses full nonlinear model which describes the evolution of a particular state of the atmosphere	Only observation operators consistent with the model	Uses simplified model describing the evolution of the mean of the pdf. Model switches can cause problems
Computational Cost	Low - local computation	Low - local computation	Medium -requires matrix inversion	Medium- high depending on number of ensemble members and resolution	Medium- Uses iterative solution methods to solve large matrix system	High
Straightforward to code and adapt to new schemes	Dependent on model parameterization - requires rewriting if a change is made	Dependent on model parameterization - requires rewriting if a change is made	Straightforward to code and adapt	Straightforward to code and adapt - although the number of ensemble members may need to change with changes in physics and resolution	Adjoint of model difficult to code but otherwise easily adaptable	Adjoint of model difficult to code but otherwise easily adaptable
Used operationally	JMA (old) mesoscale analysis in combination with OI for other variables	Met Office mesoscale analysis in conjunction with 3DVAR	NCEP Rapid Update Cycle	Never used operationally	Met Office CMC NCEP etc.	JMA ECMWF

TABLE 1. Assimilation methods and their characteristics (cont)

6. Acknowledgements

The author gratefully acknowledges Ross Bannister, Pete Clark, Mike Cullen, Mark Dixon, Jidong Gao, Brian Golding, Amos Lawless, Nancy Nichols, Ian Roulstone, Adrian Semple and Alan Thorpe for discussions and comments.

Appendix A. Sequential processing of observations with the Kalman filter

Sequential processing of observations (one-at-a-time processing) is a standard technique in discrete Kalman filtering. Here we follow Brown & Hwang (1997) is demonstrating the equivalence of sequential and simultaneous processing of observations.

We start from an alternative formulation of the Kalman filter,

$$[\mathbf{B}(t_i|y_i)]^{-1} = [\mathbf{B}(t_i)]^{-1} + [\mathbf{H}(t_i)]^T[\mathbf{R}(t_i)]^{-1}\mathbf{H}(t_i) \tag{33}$$

$$\mathbf{K}(t_i) = \mathbf{B}(t_i|y_i)\mathbf{H}(t_i)[\mathbf{R}(t_i)]^{-1}. \tag{34}$$

$$\tag{35}$$

It is easy to show via simple algebraic manipulations that this is equivalent to the filter described in section 2.1.1.

Suppose $\mathbf{R}(t_i)$ is at least block diagonal. Physically this means that the measurements available at time t_i can be grouped together in blocks (a, b, c etc. say) such that there are no correlations between measurements in different blocks. We rewrite the background error equation to make this explicit:

$$[\mathbf{B}(t_i|y_i)]^{-1} = [\mathbf{B}(t_i)]^{-1} + \left(\mathbf{H}^a(t_i)^T|\mathbf{H}^b(t_i)^T|\cdots\right) \left(\begin{array}{c|c|c} \mathbf{R}^a(t_i)^{-1} & 0 & 0 \\ \hline 0 & \mathbf{R}^b(t_i)^{-1} & 0 \\ \hline 0 & 0 & \ddots \end{array}\right) \left(\begin{array}{c} \mathbf{H}^a(t_i) \\ \hline \mathbf{H}^b(t_i) \\ \hline \vdots \end{array}\right). \tag{36}$$

Expanding the partitions of equation (36) gives

$$[\mathbf{B}(t_i|y_i)]^{-1} = \underbrace{\underbrace{[\mathbf{B}(t_i)]^{-1} + \mathbf{H}^a(t_i)^T\mathbf{R}^a(t_i)^{-1}\mathbf{H}^a(t_i)}_{[\mathbf{B}(t_i|y_i)]^{-1}\text{ after assimilating block a measurements}} + \mathbf{H}^b(t_i)^T\mathbf{R}^b(t_i)^{-1}\mathbf{H}^b(t_i) + \cdots}_{[\mathbf{B}(t_i|y_i)]^{-1}\text{after assimilating block a and b measurements}} \tag{37}$$

Note that the sum of the first two terms is just the $[\mathbf{B}(t_i|y_i)]^{-1}$ one would get by assimilating the first block of observations, just as if no further measurements were available. The Kalman gain associated with this block of measurements may now be used to update the analysis estimate accordingly. Making a trivial projection ahead through zero time allows us to choose the a posteriori $\mathbf{B}$ as the a priori $\mathbf{B}$ for the next time step. When this is added to the block b term of equation (37) we have the updated a posteriori $\mathbf{B}$ after assimilating the second block of data. This can be repeated until all observations are processed. The final estimate is then identical to that calculated had all the observations been processed simultaneously.

Now we have established the validity of sequential processing, it makes no difference whether we use the usual update formulae of section 2.1.1 or if we use the formulation given here. The final estimates will be the same (within the limits of accuracy of computational arithmetic). The choice of algorithm can be made purely on the grounds of efficiency of implementation.

Also note that, in the case where we wish to apply sequential processing but observations are mutually correlated, it is possible to take linear combinations of

observations and make new sets of measurements whose errors are uncorrelated. One way of doing this is to apply Gram-Schmidt orthogonalisation.

Appendix B. Acronyms and Abbreviations

3DVAR: Three dimensional variational assimilation
4DVAR: Four dimensional variational assimilation
AC: Analysis correction
ARPS: Advanced Regional Prediction System (University of Oklahoma)
EAKF: Ensemble adjustment Kalman filter
EnKF: Ensemble Kalman filter
FFT: Fast Fourier transform
IC: Initial condition
JMA: Japan Meteorological Agency
KF: Kalman filter
LAM: Limited area model
LBC: Lateral boundary condition
LHN: Latent Heat Nudging
MLE: Maximum likelihood estimate
NCEP: National Center for Environmental Prediction (USA)
NWP: Numerical weather prediction
OI: Optimal Interpolation
OLR: Outgoing longwave radiation
P4DVAR: Probabilistic 4DVAR
pdf: Probability density function
PI: Physical Initialisation
QPF: Quantitative precipitation forecasting
TLM: Tangent linear model
WWRP: World Weather Research Programme

References

Alberoni, P. P. 2000 Pre-processing of doppler radar data. Prepared under the European COST-717 action and available from http://pub.smhi.se/cost717/.

Alberoni, P. P., Ducrocq, V., Gregorič, G., Haase, G., Holleman, I., Lindskog, M., Macpherson, B., Nuret, M. & Rossa, A. 2000 Pre-processing of doppler radar data. Prepared under the European COST-717 action and available from http://pub.smhi.se/cost717/.

Alexander, G. D., Weinman, J. A. & Schols, J. L. 1998 The use of digital warping of microwave integrated water vapor imagery to improve forecasts of marine extratropical cyclones. *Mon. Wea. Rev.* **126**, 1469–1495.

Anderson, J. L. 2001 An ensemble adjustment Kalman filter for data assimilation. *Mon. Wea. Rev.* **129**, 2884–2903.

Aonashi, K. 1993 An initialisation method to incorporate precipitation data into a mesoscale numerical weather prediction model. *J. Meteorol. Soc. Jpn.* **71** (3), 393–406.

Bao, J.-W. & Kuo, Y.-H. 1995 On off switches in the adjoint method: step functions. *Mon. Wea. Rev.* **123**, 1589–1594.

Benjamin, S. G. 1989 An isentropic mesoα-scale analysis system and its sensitivity to aircraft and surface observations. *Mon. Wea. Rev.* **117**, 1586–1603.

Bouttier, F. & Courtier, P. 2002 *Data Assimilation Concepts and Methods*. ECMWF, meteorological Training Course Lecture Series.

Brewster, K. A. 2002 Phase-correction data assimilation and application to storm-scale numerical weather prediction. Part I: Method description and simulation testing. Submitted to Mon. Wea. Rev.

Brown, R. G. & Hwang, P. Y. C. 1997 *Introduction to random signals and applied Kalman filtering with matlab exercises and solutions*. Wiley.

Daley, R. 1991 *Atmospheric Data Analysis*. Cambridge University Press.

Derber, J. & Bouttier, F. 1999 A reformulation of the background error covariance in the ECMWF global data assimilation system. *Tellus* **51A**, 195–221.

Ducrocq, V., Gregorič, G., Haase, G., Macpherson, B., Nuret, M. & Rossa, A. 2000 Assimilation of radar precipitation data in NWP models - a review. Prepared under the European COST-717 action and available from http://pub.smhi.se/cost717/.

Ebert, E. 2002 Verification webpage. http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html, published under the auspices of the WWRP Verification Group.

Gao, J., Xue, M., Brewster, K., Carr, F. & Drogemeier, K. K. 2002 New developments of a 3DVAR system for a nonhydrostatic NWP model. In *Preprint, 15th Conf. on Numerical Weather Prediction and 19th Conf. on Weather Analysis and Forecasting*. AMS, San Antonio, Texas.

Gauthier, P., C. Charrette, L. F., Koclas, P. & Laroche, S. 1999 Implementation of a 3D variational data assimilation at the Canadian Meteorological Centre. Part I: The global analysis. *Atmosphere-Ocean* **37** (2), 103–156.

Golding, B. W. 1987 Strategies for using mesoscale data in an operational mesoscale model. In *Proc. Symp. Mesoscale Analysis and Forecasting*. Vancouver, Canada.

Gregory, D. & Rowntree, P. R. 1990 A mass flux convection scheme with representation of cloud ensemble characteristics and stability-dependent closure. *Mon. Wea. Rev.* **118**, 1483–1506.

Gustafsson, N., Berre, L., Hörnquist, S., Huang, X.-Y., Lindskog, M., Navascués, B., Mogensen, K. S. & Thorsteinsson., S. 2001 Three-dimensional variational data assimilation for a limited area model. *Tellus* **53A** (4).

Hamill, T. M. 2002 Analysis-error statistics of a quasigeostrophic model using three-dimensional variational assimilation. *Mon. Wea. Rev.* **130**, 2777–2790.

Hollingsworth, A. & Lonnberg, P. 1986 The statistical structure of short-range forecast errors as determined by radiosonde data. Part I: The wind field. *Tellus* **38A**, 111–136.

Houtekamer, P. L. & Mitchell, H. L. 2001 A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.* **129**, 123–137.

Huang, X.-Y. 2000 Variational analysis using spatial filters. *Mon. Wea. Rev.* **128**, 2588–2600.

Ide, K., Courtier, P., Ghil, M. & Lorenc, A. C. 1997 Unified notation for data assimilation: Operational, sequential and variational. *J. Met. Soc. Japan* **75** (1B), 181–189.

Ingleby, N. B. 2002 Options for improving the modelling of background errors in the Met Office variational analysis. Unpublished.

Jazwinski, A. H. 1970 *Stochastic Processes and Filtering Theory*. Academic Press.

Jones, C. D. & Macpherson, B. 1997 A latent heat nudging scheme for the assimilation of precipitation data into an operational mesoscale model. *Meteorol. Appl.* pp. 269–277.

Krishnamurti, T. N., Xue, J. & Bedi, H. S. 1991 Physical initialization for numerical weather prediction over the tropics. *Tellus* **43** (4), 53–81.

Kuizumi, K., Ishikawa, Y., Tsuyuki, T. & Murakami, S. 2002 Improvement in precipitation forecasts of the JMA mesoscale model with four-dimensional variational data assimlation. Presentation at the WWRP International Conference of Quantitative Precipitation Forecasting, University of Reading, UK.

Lermusiaux, P. F. J. & Robinson, A. R. 1999 Data assimilation via error subspace statistical estimation. *Mon. Wea. Rev.* **127**, 1385–1407.

Lorenc, A. C. 1981 A global three dimensional multivariate statistical interpolation scheme. *Mon. Wea. Rev.* **109**, 701–721.

Lorenc, A. C. 1986 Analysis methods for numerical weather prediction. *Q. J. R. Meteorol. Soc.* **112**, 1177–1194.

Lorenc, A. C. 1997 Development of an operational variational assimilation scheme. *J. Meteorol. Soc. Jpn.* **75** (1B), 339–346.

Lorenc, A. C. 2002*a* Modelling of error covariances by four-dimensional variational data assimilation. Forecasting Research Scientific Paper 68. Met Office, submitted to Q. J. R. Meteorol. Soc.

Lorenc, A. C. 2002*b* Modelling of error covariances by the ensemble Kalman filter. Forecasting Research Scientific Paper 69. Met Office, submitted to Q. J. R. Meteorol. Soc.

Lorenc, A. C., Ballard, S. P., Bell, R. S., Ingelby, N. B., Andrews, P. L. F., Barker, D. M., Bray, J. R., Clayton, A. M., Dalby, T., Li, D., Payne, T. J. & Saunders, F. W. 2000 The Met Office global three-dimensional variational data assimilation scheme. *Q. J. R. Meteorol. Soc.* **126** (570), 2991–3012.

Mitchell, H. L., Houtekamer, P. L. & Pellerin, G. 2002 Ensemble size, balance, model-error representation in an ensemble Kalman filter. *Mon. Wea. Rev.* **130**, 2791–2808.

Parrish, D. F. & Derber, J. C. 1992 The National Meteorological Center's spectral statistical-interpolation analysis scheme. *Mon. Wea. Rev.* **120**, 1747–1763.

Raymond, W. H. & Garder, A. 1991 A review of recursive and implicit filters. *Mon. Wea. Rev.* pp. 477–495.

Shin, D. W. & Ahn, J. B. 2002 Adaptive use of TRMM observations for tropical precipitation forecasts. *J. Met. Soc. Japan* **60** (1), 85–97.

Smith, L. A., Ziehmann, C. & Fraedrich, K. 1999 Uncertainty dynamics and predictability in chaotic systems. *Q. J. R. Meteorol. Soc.* **125** (560), 2855–2866.

Sun, J. & Crook, N. A. 2001 Real-time low-level wind and temperature analysis using single wsr-88d data. *Weather Forecast.* **16**, 117–132.

Tsuyuki, T. & Fujita, T., ed. 2002 *Outline of the operational numerical weather prediction at the Japan Meteorological Agency*. Japan Meteorological Agency, appendix to WMO numerical weather prediction progress report.

Wang, W. & Warner, T. T. 1988 Use of four-dimensional data assimilation by Newtonian and latent-heat forcing to improve a mesoscale-model precipitation forecast: a case study. *Mon. Wea. Rev.* pp. 2593–2613.

Weygandt, S. S., Shapiro, A. & Droegemeier, K. K. 2002*a* Retrieval of model initial fields from single-doppler observations of a supercell thunderstorm. Part I: Single-doppler velocity retrieval. *Mon. Wea. Rev.* **130**, 433–453.

Weygandt, S. S., Shapiro, A. & Droegemeier, K. K. 2002*b* Retrieval of model initial fields from single-doppler observations of a supercell thunderstorm. Part II: Thermodynamic retrieval and numerical prediction. *Mon. Wea. Rev.* **130**, 454–476.

Wu, W.-S. & Purser, R. J. 2002 Three-dimensional variational analysis with spatially inhomogeneous covariances. *Mon. Wea. Rev.* **130**, 2905–2916.

Xie, Y., Lu, C. & Browning, G. L. 2002 Impact of formulation of cost function and constraints on three-dimensional variational data assimilation. *Mon. Wea. Rev.* **130**, 2433–2447.

Xue, M., Wang, D., Gao, J., Brewster, K. & Droegemeier, K. K. 2002 The Advanced Regional Prediction system (ARPS), storm-scale numerical weather prediction and data assimilation. *Meteorol. Atmos. Phys.* To appear.

Zupanski, D. 1997 A general weak constraint approach to operational 4DVAR data assimilation systems. *Mon. Wea. Rev.* **125**, 2274–2291.